Analysis of Factors Influencing Income in Canada

Myisha Chaudhry - 200591740
Kayleigh Habib - 200370580
Abigail Lee - 200469770

ST362: Regression Analysis
Wilfrid Laurier University
Dr. Robert Rundle
July 25, 2022

Table of Contents

Introduction

This report will investigate some factors that influence average employment income for people residing in Canada. The data which is titled: "Immigrant Status and Period of Immigration (10), Employment Income Statistics (7), STEM and BHASE (non-STEM) Groupings, Major Field of Study - Classification of Instructional Programs (CIP) 2016 (36), Highest Certificate, Diploma or Degree (9), Work Activity During the Reference Year (3), Age (10) and Sex (3) for the Population Aged 15 Years and Over in Private Households of Canada, Provinces and Territories and Census Metropolitan Areas, 2016 Census - 25% Sample Data" is gathered from the 2016 census and accessed through the Statistics Canada website. More detailed data was not yet available for the recent 2021 census; thus the 2016 census data was used in this analysis.

The main question we will be answering through this analysis is: What are the factors that impact income levels in Canada? Some sub questions that we will also be working to answer include the following:

- Does gender affect salary?
- Does province of residence affect salary?
- Does citizenship/immigration status affect salary?
- Does education level affect salary?

We believe that all of these factors do have some level of impact in the average salary of Canadians. The following are our hypotheses:

1. Given our knowledge of the gender pay gap, we believe that males will be earning more than their female counterparts.
2. We suspect that those residing in more populated provinces (such as Ontario or British Columbia) with numerous metropolitan areas will have higher salaries than those residing in the other provinces.
3. We predict that immigrants will earn less than non-immigrants.
4. Lastly, we think that people with higher education will earn more than those with lower or none at all.

Data Description

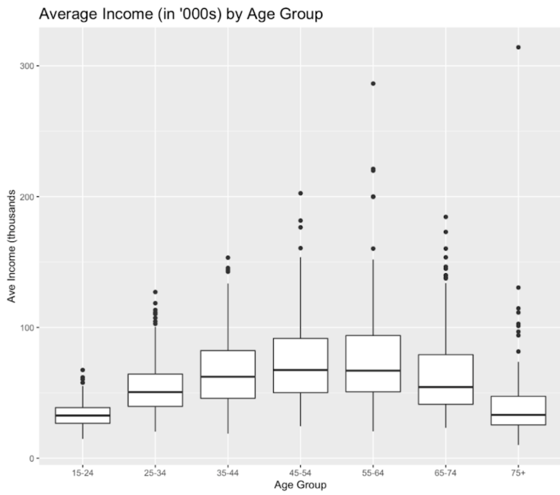
The dataset contains over 14 million records in 32 columns. As this included more details than required, we performed appropriate filtering, column selection and renaming of columns to make the data more meaningful. The reduced dataset was used for the remainder of this project contains the following columns:

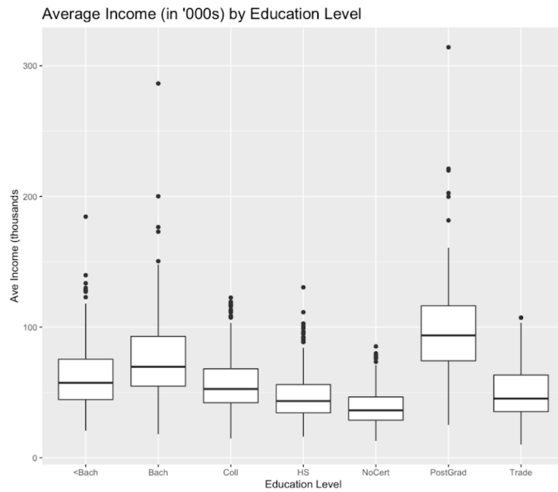
Variable	Description
Age	7 factor levels : age groups 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+
EducInd	7 factor levels :No certificate, High School, Trade, College, <Bachelor, Bachelor, Post Graduate
Gender	Male or Female
Immigration Status	Immigrant or Non-Immigrant
Province	AB, BC, MN, NB, NFLD, NS, ON, PEI, QC, SK, NT, NWT, YK
Income (Target)	Average Income in in thousands

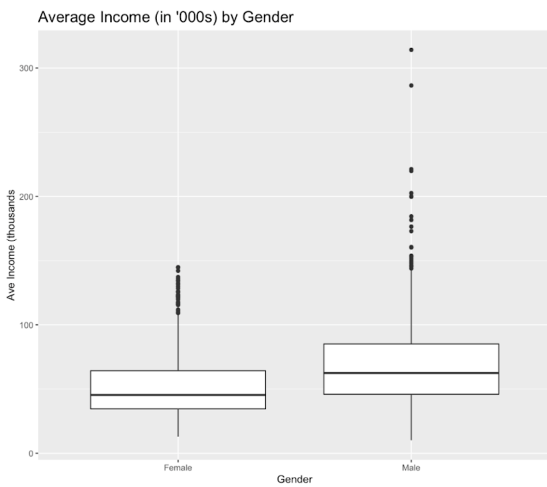
```
     Age          Gender           EducInd                  ImmigStatus        Province     AverageIncome
 15-24:236   Female:910   <Bach      :224   Immigrants     : 777   ON       :193   Min.   : 10.14
 25-34:307   Male  :972   Bach       :289   Non-immigrants:1105   BC       :185   1st Qu.: 38.98
 35-44:329                Coll       :298                         QC       :183   Median : 54.22
 45-54:337                HS         :291                         AB       :182   Mean   : 61.43
 55-64:324                NoCert     :255                         MN       :171   3rd Qu.: 76.87
 65-74:259                PostGrad:271                            SK       :164   Max.   :314.18
 75+  : 90                Trade      :254                         (Other):804
```

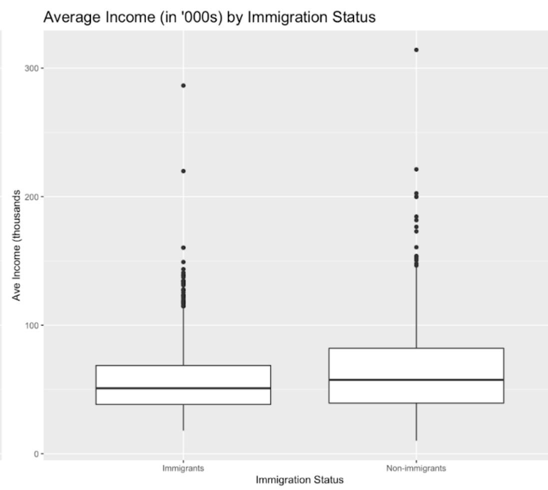
Exploratory Data Analysis

To begin our analysis, we created boxplots to display the relationship of each of the predictors to the target variable (income) in the data.



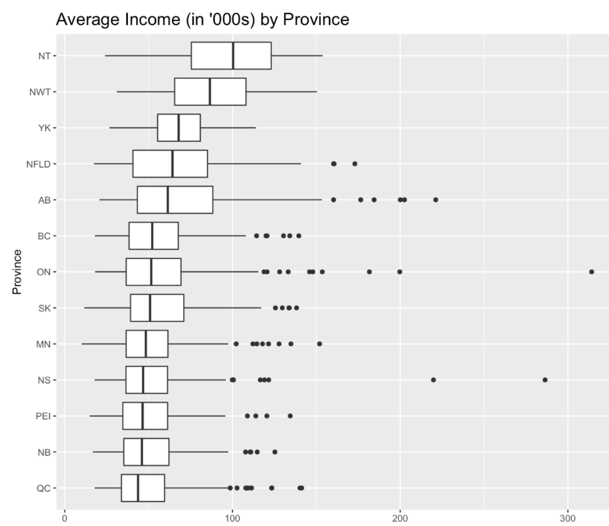






Based on these charts, we can observe the following:

1. Average employment income increases with age up to the 55-64 age group. After that point it declines. This is likely because many people retire between the ages of 60 to 65.
2. Average employment income increases as education level increases.
3. Average employment income is higher for males than females.
4. There are some provinces that have very similar distributions of Average Income. Nunavut appears to have the highest median value.
5. While non-immigrants seem to have slightly higher incomes, this difference is small.



Model 1 (Age + Province + Gender + EducInd + ImmigStatus)

For the first model, we have chosen to use all the predictors in the data. Since the predictors are categorical variables, there are numerous coefficients based on the level of the category.

```
Call:
lm(formula = AverageIncome ~ Age + Province + Gender + EducInd +
    ImmigStatus, data = projectData)

Residuals:
    Min      1Q  Median      3Q     Max
-53.440  -7.624  -1.157   5.923 221.854

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                33.4997     1.8668  17.945  < 2e-16 ***
Age25-34                   16.8346     1.3477  12.491  < 2e-16 ***
Age35-44                   29.2299     1.3330  21.928  < 2e-16 ***
Age45-54                   37.3022     1.3281  28.086  < 2e-16 ***
Age55-64                   38.8530     1.3357  29.087  < 2e-16 ***
Age65-74                   27.4578     1.3953  19.679  < 2e-16 ***
Age75+                      9.3121     1.9428   4.793 1.77e-06 ***
ProvinceBC                -13.6945     1.6161  -8.474  < 2e-16 ***
ProvinceMN                -18.0859     1.6499 -10.962  < 2e-16 ***
ProvinceNB                -23.6351     1.7408 -13.577  < 2e-16 ***
ProvinceNFLD               -9.7569     1.8652  -5.231 1.88e-07 ***
ProvinceNS                -20.5264     1.7089 -12.012  < 2e-16 ***
ProvinceNT                 18.6299     2.0360   9.150  < 2e-16 ***
ProvinceNWT                10.5843     1.8764   5.641 1.95e-08 ***
ProvinceON                -10.6797     1.6019  -6.667 3.44e-11 ***
ProvincePEI               -27.3892     1.9231 -14.242  < 2e-16 ***
ProvinceQC                -19.9015     1.6203 -12.283  < 2e-16 ***
ProvinceSK                -14.2393     1.6685  -8.534  < 2e-16 ***
ProvinceYK                 -9.3225     1.9976  -4.667 3.28e-06 ***
GenderMale                 18.1864     0.7160  25.401  < 2e-16 ***
EducIndBach               14.0044     1.3862  10.103  < 2e-16 ***
EducIndColl                -5.0241     1.3781  -3.646 0.000274 ***
EducIndHS                 -13.5872     1.3837  -9.819  < 2e-16 ***
EducIndNoCert             -23.4217     1.4210 -16.482  < 2e-16 ***
EducIndPostGrad            32.1177     1.4044  22.870  < 2e-16 ***
EducIndTrade              -12.3858     1.4214  -8.714  < 2e-16 ***
ImmigStatusNon-immigrants   9.8937     0.7497  13.196  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.48 on 1855 degrees of freedom
Multiple R-squared:  0.7539,    Adjusted R-squared:  0.7505
F-statistic: 218.6 on 26 and 1855 DF,  p-value: < 2.2e-16
```

The model is the following:

Y = B0 + B1*(Age) + B2*(Province) + B3*(Gender) + B4*(Education Level) + B5*(ImmigStatus) + error

```
Analysis of Variance Table

Response: AverageIncome
              Df Sum Sq Mean Sq F value    Pr(>F)    
Age            6 350446   58408  243.84 < 2.2e-16 ***
Province      12 274107   22842   95.36 < 2.2e-16 ***
Gender         1 147260  147260  614.77 < 2.2e-16 ***
EducInd        6 547699   91283  381.08 < 2.2e-16 ***
ImmigStatus    1  41712   41712  174.14 < 2.2e-16 ***
Residuals   1855 444340     240                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the ANOVA table above, it appears that all predictors are significant in this model. The p-values for all predictors are less than 0.05 (alpha), meaning that we reject the null hypothesis that these predictors are not significant.

Model 2 (Age (<65 only) + Province + Gender + EducInd + ImmigStatus)

Based on our data visualization, we observed that average income increases with age up to 64 and declines thereafter. This is likely due to the fact that most Canadians are retired by age 65. We decided to remove the highest two age groups from the data, i.e. ages 65-74 and ages 75+.

```
   Age          Gender          EducInd             ImmigStatus        Province    AverageIncome
 15-24:236   Female:756   <Bach    :188   Immigrants    :653   AB     :140   Min.   : 14.89
 25-34:307   Male  :777   Bach     :242   Non-immigrants:880   BC     :140   1st Qu.: 39.65
 35-44:329                Coll     :241                        ON     :140   Median : 55.24
 45-54:337                HS       :231                        QC     :140   Mean   : 62.17
 55-64:324                NoCert   :202                        MN     :136   3rd Qu.: 77.32
                          PostGrad:222                         SK     :132   Max.   :286.41
                          Trade    :207                        (Other):705
```

Y = B0 + B1*(Age < 65) + B2*(Province) + B3*(Gender) + B4*(Education Level) + B5*(ImmigStatus) + error

```
Residual standard error: 13.74 on 1508 degrees of freedom
Multiple R-squared:  0.7973,    Adjusted R-squared:  0.7941
F-statistic: 247.1 on 24 and 1508 DF,  p-value: < 2.2e-16
```

We looked at the summary function on this model and compared to our first model. Comparing the adjusted R^2 for this model (0.7941) with that for the original model (0.7505), we see that the updated model explains more of the variation in the average income. In this case, approximately 80% versus 75% for the original model. Therefore, we will use this updated model for the next part of the analysis.

Model 3 (Age + Province + Gender + EducInd)

Y = B0 + B1*(Age) + B2*(Province) + B3*(Gender) + B4*(Education Level) + error

We can test a simplified model that omits one or more of the predictors. In the following model, we omitted immigration status since from our data exploration, there did not appear to be a material difference in income levels by this factor. We then tested this simplified model against the full model.

Here, we can do the following hypothesis test:

$H_{0:}$ the coefficients for the variables in the full model which are not in the simplified model, are zero.

$H_{a:}$ the coefficients are not zero, meaning that the full model is better.

```
Analysis of Variance Table

Model 1: AverageIncome ~ Age + Province + Gender + EducInd
Model 2: AverageIncome ~ Age + Province + Gender + EducInd + ImmigStatus
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1509 327959
2   1508 284784  1     43175 228.62 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the ANOVA test, since the p-value is less than 0.05, we will reject the null hypothesis. Thus we would conclude that the full model is better than the simplified model.

<u>Dummy Variables</u>

As we are using categorical variables in our analysis, we ensured that we knew which level of each variable was set as the base level. For each of the categorical variables, the following show the dummy variables being assigned:

	25-34	35-44	45-54	55-64
15-24	0	0	0	0
25-34	1	0	0	0
35-44	0	1	0	0
45-54	0	0	1	0
55-64	0	0	0	1

	BC	MN	NB	NFLD	NS	NT	NWT	ON	PEI	QC	SK	YK
AB	0	0	0	0	0	0	0	0	0	0	0	0
BC	1	0	0	0	0	0	0	0	0	0	0	0
MN	0	1	0	0	0	0	0	0	0	0	0	0
NB	0	0	1	0	0	0	0	0	0	0	0	0
NFLD	0	0	0	1	0	0	0	0	0	0	0	0
NS	0	0	0	0	1	0	0	0	0	0	0	0
NT	0	0	0	0	0	1	0	0	0	0	0	0
NWT	0	0	0	0	0	0	1	0	0	0	0	0
ON	0	0	0	0	0	0	0	1	0	0	0	0
PEI	0	0	0	0	0	0	0	0	1	0	0	0
QC	0	0	0	0	0	0	0	0	0	1	0	0
SK	0	0	0	0	0	0	0	0	0	0	1	0
YK	0	0	0	0	0	0	0	0	0	0	0	1

	Male
Female	0
Male	1

	Non-immigrants
Immigrants	0
Non-immigrants	1

	<Bach	Bach	Coll	HS	PostGrad	Trade
NoCert	0	0	0	0	0	0
<Bach	1	0	0	0	0	0
Bach	0	1	0	0	0	0
Coll	0	0	1	0	0	0
HS	0	0	0	1	0	0
PostGrad	0	0	0	0	1	0
Trade	0	0	0	0	0	1

Based on the above, we note the following:

1. The 15-24 age group is the base level for Age
2. Alberta (AB) is the base level for Province
3. Females are the base level for Gender
4. No certificate is the base level for Education
5. Immigrant is the based level for Immigration Status

Analysis of Residuals

In linear models, there are several assumptions that apply to the residuals:

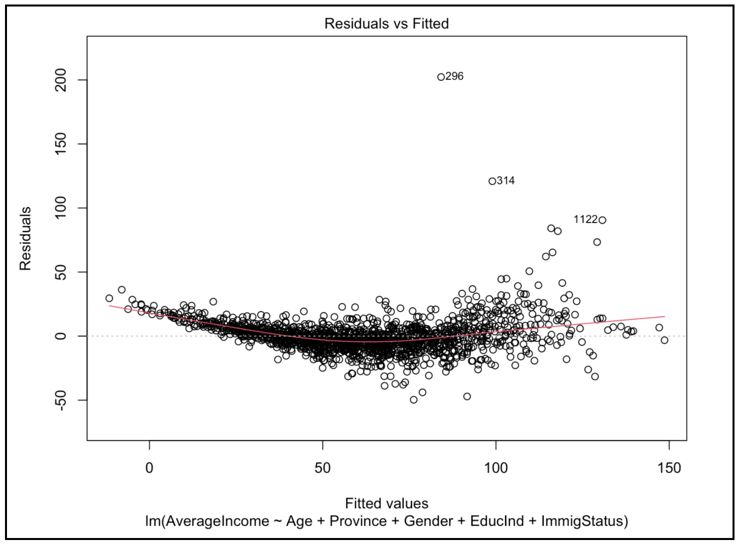
a. residuals have a mean of zero
b. residuals are normally distributed
c. residuals are homoscedastic (equal variances)
d. residuals are independent

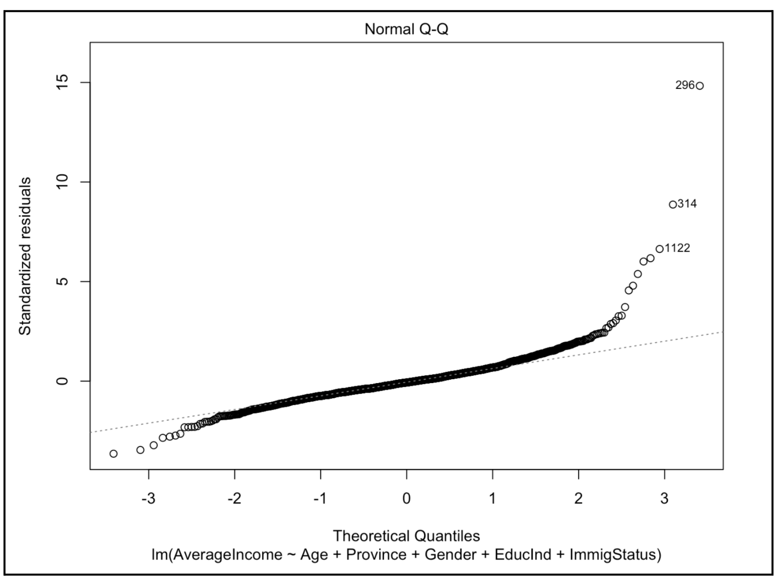
We used the plot() function in r to look at the residuals from the model:

Plots

Plot 1

This plot checks the linear relationship assumptions and a horizontal line without any patterns is an indication for a good linear relationship. In the case of the plot we created for model 3, the points do not follow a straight line, and the distribution around is not even. Thus there may be some patterns in the residuals, so the assumption of linearity is not being met.



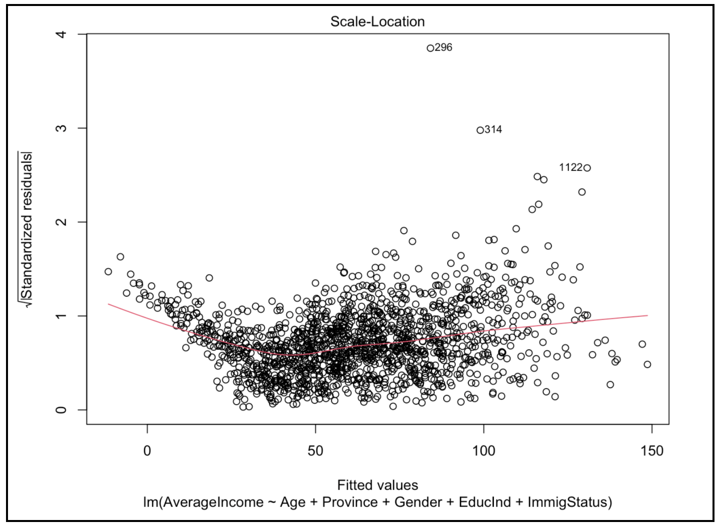


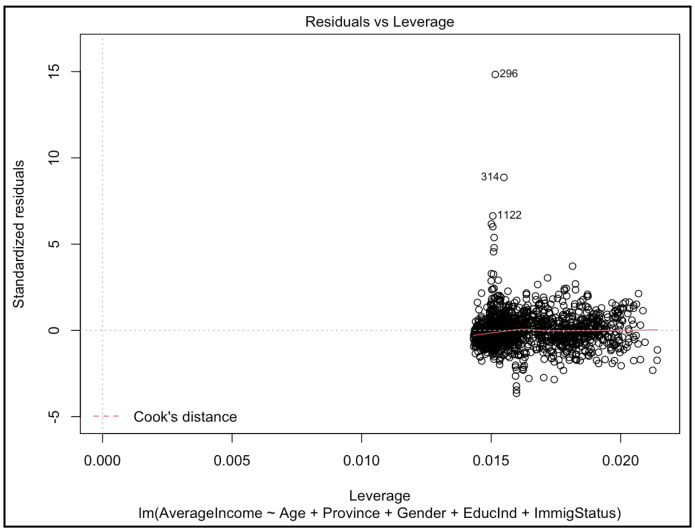
Plot 2

This plot is a Normal Q-Q plot used to examine whether the residuals are normally distributed. The plot created for model 3 has the majority of the points following the dashed line meaning most of them follow the normal distribution. Around either end of the line the data deviates and no longer follows the straight dashed line. This tells us that these end data points do not follow a normal distribution.

Plot 3

This plot is a Scale-Location plot of the square root of residuals against fitted values, used to check the homogeneity of variance of the residuals (homoscedasticity). The plot created for model 3 suggests some non-linearity, but the spread of the magnitudes seems to be lowest in the fitted values between 100 and 150, the spread of the magnitudes is highest in the fitted values between 25 and 75, and medium around 0.



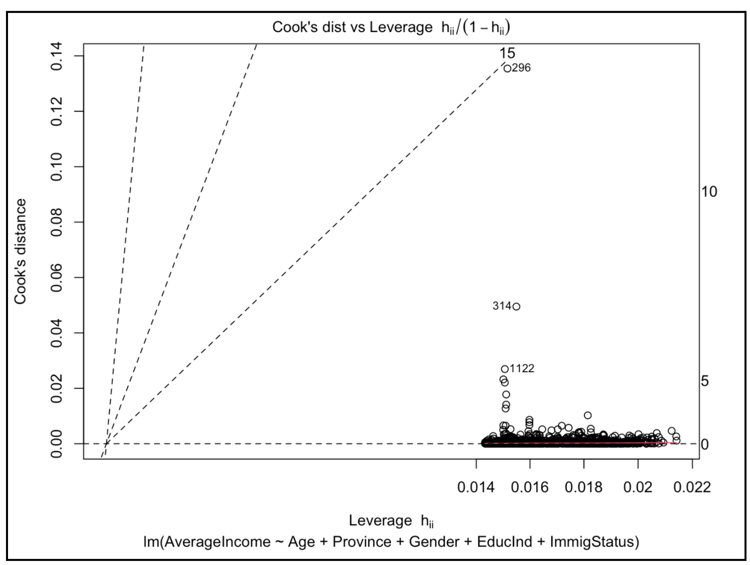


Plot 5

This plot is a plot of residuals against leverages and used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. Based on this plot, it does appear that there are some cases that do not fall within the range of other data points.

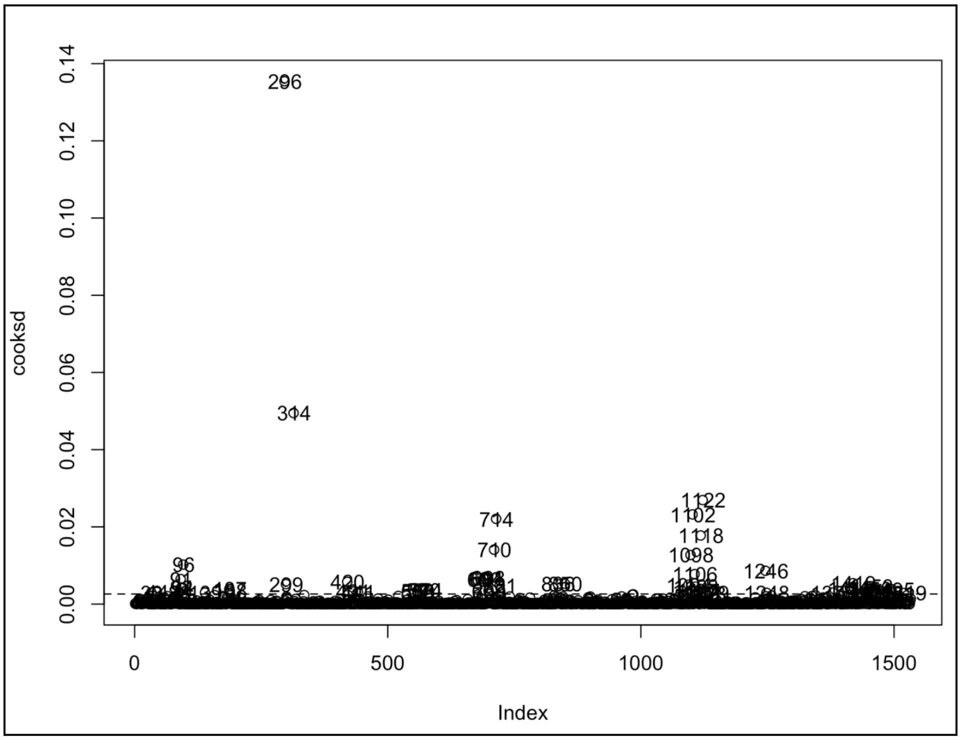
Plot 6

This is a plot of Cook's distances against leverage/(1-leverage). For plot 6, we can see that there is a red dashed line near the bottom. This line tells us that the observations are all not inside of Cook’s distance, meaning that there are some influential observations. This essentially tells us that in our data, there are some outliers present.



Cook's Distance

To further our statistical understanding of our data we decided to calculate Cook's Distance. From this plot it is evident that there are a few data points that do not necessarily follow the data and are considered influential (or outliers). We also identified the true numerical values of these influential points and there are a total of 55 of them. We want to remove these influential values from the model as they create a certain bias in our data.




```
 [1]   33   44   88   91   92   94   96  139  187  188  296  299  314  420  431  441  558  559
570  574  690  694
[23]  698  699  700  710  714  721  836  850 1094 1098 1102 1106 1107 1108 1109 1114 1118 1122
1123 1129 1246 1248
[45] 1379 1413 1419 1430 1450 1453 1455 1459 1473 1495 1519
```

Model 4 (Model 2 without the influential points)

To better understand what our data is trying to tell us, removing the influential values can give a clearer indication of the trends that are present.

```
Residual standard error: 9.541 on 1453 degrees of freedom
Multiple R-squared:  0.8762,    Adjusted R-squared:  0.8741
F-statistic: 428.3 on 24 and 1453 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: AverageIncome
              Df Sum Sq Mean Sq F value    Pr(>F)    
Age            4 250361   62590  687.54 < 2.2e-16 ***
Province      12 237735   19811  217.62 < 2.2e-16 ***
Gender         1  83012   83012  911.86 < 2.2e-16 ***
EducInd        6 323705   53951  592.63 < 2.2e-16 ***
ImmigStatus    1  40994   40994  450.30 < 2.2e-16 ***
Residuals   1453 132274      91                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the ANOVA test for this model, we get a very low p-value which means that we reject the null hypothesis and thus this model is better than model 2. We can also compare the adjusted R^2 values from model 2 (0.7941) and model 4 (0.8741) and see that this model has more predictive power since the adjusted R^2 value is much higher than that of model 2.

<u>Test for Homoscedasticity</u>

For this test the hypotheses are as follows:

> H_o: the error variances are all equal
> H_a: the error variances are not equal, i.e. as average income increases, the variances increase (or decrease).

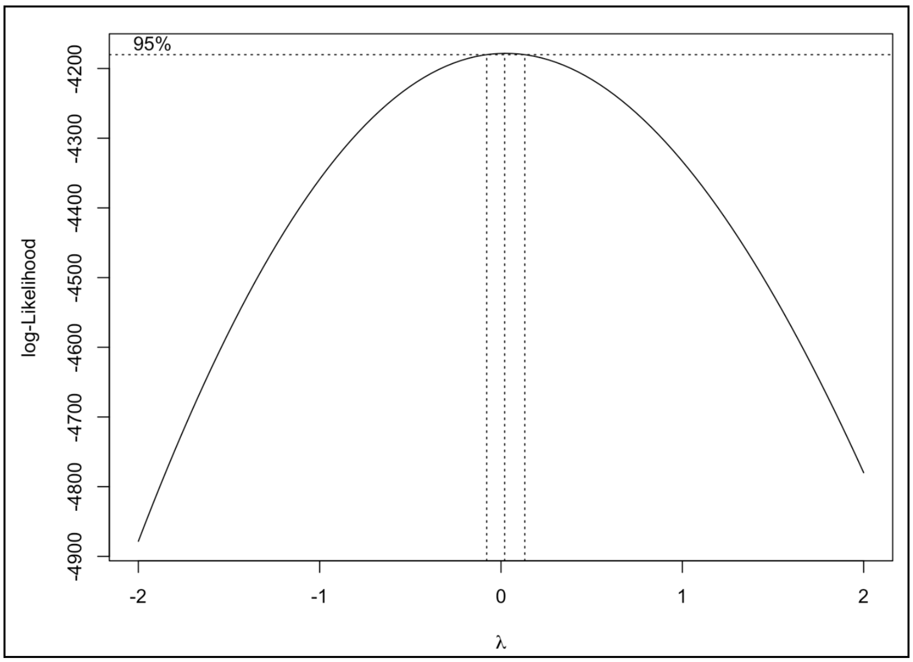
````
```{r}
bptest(model4)
```

	studentized Breusch-Pagan test

data:  model4
BP = 238.67, df = 24, p-value < 2.2e-16
````

If heteroscedasticity is present, meaning that the error terms are not equally scattered, then we have to apply a transformation. By applying the Breusch-Pagan test to determine if heteroscedasticity is present, we obtained a p-value less than 0.05. Thus we reject the null hypothesis and conclude that there is sufficient evidence to say that heteroscedasticity is present in the regression model. Thus, the standard errors that are shown in the output table of the regression may be unreliable.

Since heteroscedasticity is present, we apply an appropriate transformation.

<u>Box-Cox Analysis</u>

The box-cox analysis indicates that a transformation may be useful for the response and regressor variables. When doing the box-cox analysis for our model, it is evident that in the plot, when lambda is 0 it falls into the 95% confidence interval. This means that the kind of transformation that would be needed to make our model better is the log transformation.



Model 5 (Model 4 with log transformation)

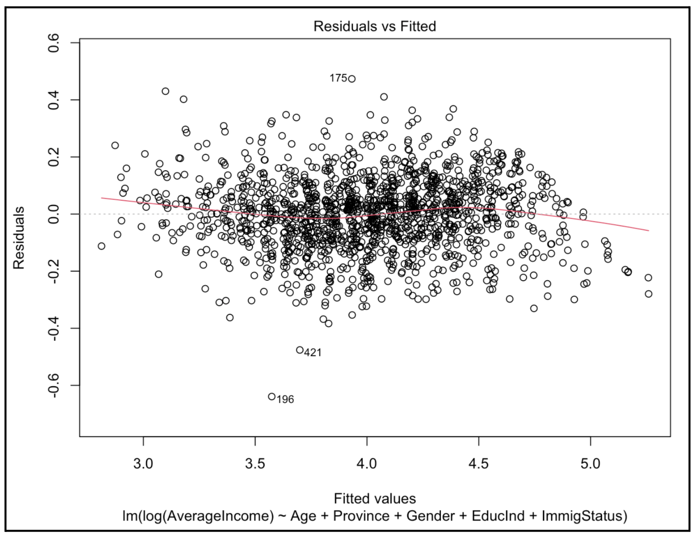
The purpose of the log transformation is to allow our model to become more predictive. After performing the log transformation on our model, the adjusted R^2 value is now 0.9127, meaning that our model has improved as prior to this transformation the adjusted R^2 value was 0.8741.

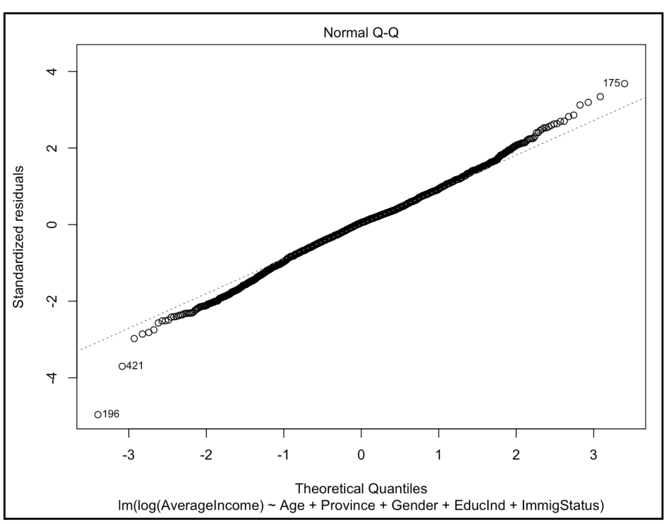
```
Residual standard error: 0.1299 on 1453 degrees of freedom
Multiple R-squared:  0.9141,    Adjusted R-squared:  0.9127
F-statistic: 644.6 on 24 and 1453 DF,  p-value: < 2.2e-16
```

To verify that this model meets the conditions required for linear regression, better we used the plot() function again.

<u>Plots</u>

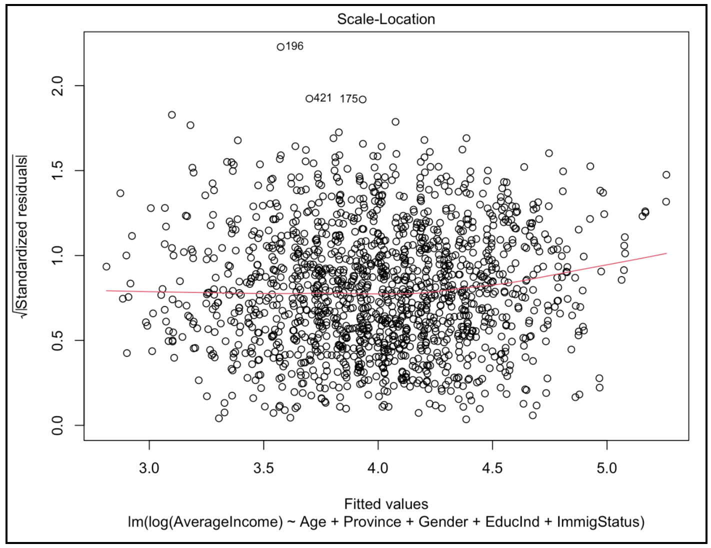
Plot 1
We see that in this plot the red line is almost horizontal, and the residuals are evenly distributed about this line. This is an indication of a mean close to zero.

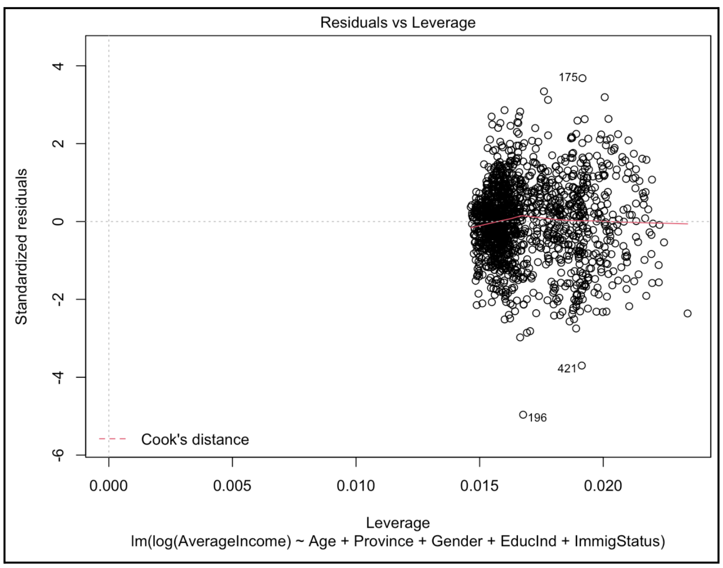




Plot 2
We see in this plot that the majority of the points lie on the straight dashed line. This means that the residuals follow a normal distribution.

Plot 3
These points are fairly sparse and are evenly distributed about the red line, which indicates that there is homogeneity within the data.





Plot 5
This plot indicates that there are no highly influential cases remaining in the data.

Based on these plots, it now appears that the assumptions for linear regression are met. In addition, our model has strong predictive power as indicated by the high adjusted R^2 value, so that the predictor variables included explain 91% of the variation in income.

Conclusion

Our analysis using regression models and statistical methods allowed us to determine that age group, province, gender, education level and immigration status all have an impact on income levels in Canada.

After visualizing our data, we used an ANOVA test to compare a model with all predictors versus a simplified model that omitted immigration status. This test indicated that the full model was better. We then removed 2 age levels that represented post-retirement individuals so that we would focus primarily on employment earnings. We then analyzed the residuals to determine whether they were consistent with the assumptions for linear models. Based on this we removed some influential cases to reduce bias in our model, and we applied a log transformation to the target variable based on the results of a Box-Cox analysis. The final model is shown as follows:

```
Call:
lm(formula = log(AverageIncome) ~ Age + Province + Gender + EducInd +
    ImmigStatus, data = newProjDataAdj)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63911 -0.07768  0.00649  0.07935  0.47345

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.072991   0.017162 179.057  < 2e-16 ***
Age25-34                   0.376804   0.011657  32.325  < 2e-16 ***
Age35-44                   0.564366   0.011549  48.867  < 2e-16 ***
Age45-54                   0.667739   0.011548  57.823  < 2e-16 ***
Age55-64                   0.668895   0.011626  57.533  < 2e-16 ***
ProvinceBC                -0.205236   0.015943 -12.873  < 2e-16 ***
ProvinceMN                -0.265923   0.016058 -16.560  < 2e-16 ***
ProvinceNB                -0.372874   0.016551 -22.529  < 2e-16 ***
ProvinceNFLD              -0.189018   0.017924 -10.545  < 2e-16 ***
ProvinceNS                -0.355029   0.016398 -21.651  < 2e-16 ***
ProvinceNT                 0.226352   0.019094  11.855  < 2e-16 ***
ProvinceNWT                0.135140   0.017682   7.643 3.84e-14 ***
ProvinceON                -0.199442   0.016051 -12.426  < 2e-16 ***
ProvincePEI               -0.444463   0.018053 -24.620  < 2e-16 ***
ProvinceQC                -0.348305   0.015995 -21.775  < 2e-16 ***
ProvinceSK                -0.171208   0.016093 -10.639  < 2e-16 ***
ProvinceYK                -0.106112   0.018187  -5.835 6.64e-09 ***
GenderMale                 0.291754   0.006772  43.085  < 2e-16 ***
EducInd<Bach               0.455125   0.013326  34.153  < 2e-16 ***
EducIndBach                0.634846   0.012651  50.180  < 2e-16 ***
EducIndColl                0.381154   0.012569  30.324  < 2e-16 ***
EducIndHS                  0.199707   0.012657  15.779  < 2e-16 ***
EducIndPostGrad            0.815238   0.013388  60.893  < 2e-16 ***
EducIndTrade               0.260382   0.012981  20.058  < 2e-16 ***
ImmigStatusNon-immigrants  0.184258   0.007054  26.119  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the coefficients indicated for our final model, it is evident that:

1. Income levels increase as age increases above the base level of 15-24. This is consistent with our hypothesis.
2. Using Alberta as the base level, some regions (Nunavut and Northwest Territories) have a positive impact on income, while the remaining regions have a negative impact. This was different from what we had originally thought. According to Statistics Canada, this is a consequence of the economic “boom” in the resource-based sector, particularly in Nunavut and Northwest Territories, that coincided with a decline in the manufacturing sector which impacted the economies of Ontario and Québec.
3. Males have higher incomes than females, this is consistent with our hypothesis.
4. Education level also impacts income. For each higher level of education, income levels also increase. This is also consistent with our expectation.
5. Non-immigrant status also contributes in a positive way to income. This is consistent with our hypothesis.

This type of analysis can be useful to government policy-makers, economists, as well as academic planners. Some work has been done to reduce the gender pay-gap, but our analysis shows that there is still room for increased equity in income levels between males and females. The income differences observed by education level can be used to implement programs aiming to improve access to higher education, regardless of immigration status. Finally, the regional differences in income help highlight the need for greater diversity in economies so that there is less variation based on residence.

Works Cited

Government of Canada, Statistics Canada. “Data Tables, 2016 Census.” *Immigrant Status and Period of Immigration (10), Employment Income Statistics (7), STEM and BHASE (Non-STEM) Groupings, Major Field of Study - Classification of Instructional Programs (CIP) 2016 (36), Highest Certificate, Diploma or Degree (9), Work Activity During the Reference Year (3), Age (10) and Sex (3) for the Population Aged 15 Years and Over in Private Households of Canada, Provinces and Territories and Census Metropolitan Areas, 2016 Census - 25% Sample Data*, 17 June 2019, https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dt-td/Rp-eng.cfm?TABID=1&LANG=E&A=R&APATH=3&DETAIL=0&DIM=0&FL=A&FREE=0&GC=01&GL=-1&GID=1325190&GK=1&GRP=1&O=D&PID=110936&PRID=10&PTYPE=109445&S=0&SHOWALL=0&SUB=0&Temporal=2017&THEME=123&VID=0&VNAMEE=&VNAMEF=&D1=0&D2=0&D3=0&D4=0&D5=0&D6=0.

Government of Canada, Statistics Canada. “Household Income in Canada: Key Results from the 2016 Census.” *The Daily* - , 27 Sept. 2017, https://www150.statcan.gc.ca/n1/daily-quotidien/170913/dq170913a-eng.htm.